MAS223 Statistical Inference and Modelling

Dr Nic Freeman

December 7, 2015

Contents

Chapter 1

Univariate Distribution Theory

1.1 Sample spaces, events and random variables

We start with some revision of material from first-year courses, in particular MAS113 Introduction to Probability and Statistics.

In probability and statistics we are usually interested in situations (often referred to as **experiments**) where we have some uncertainty about the outcome. In any such experiment we are able to identify a set S of possible outcomes, known as the **sample space**; one and only one of these outcomes will actually be observed when the experiment is performed.

Events are subsets of the sample space S. If $A \subseteq S$ is an event then the observed outcome may or may not be a member of A; if it is, we say that A **occurs**. To every event we associate a **probability**, $P(A)$, which we think of as the chance of the event A occurring.

Probabilities obey the **axioms of probability** (see MAS113) and the many theorems that follow from these axioms. This allows us to deduce the probability of (possibly complicated) events, for example by considerations of symmetry or through approximation by simpler events. Often, we already know the probabilities of some events, and we need to think about how to find out the probabilities of other events.

Frequently, we are interested in a numerical measurement arising from an experiment, rather than the raw outcome - for example, we might count the number of heads in a sequence of coin tosses, rather than recording the exact sequence of heads and tails. In such situations we work with a **random variable** X, defined as a real function $X : S \to \mathbb{R}$. Then, X associates each element of the sample space to a real number, and we are interested in probabilities of the form $P(X \in E)$, where E is a subset of $\mathbb{R}$. These probabilities form the **distribution** (or **probability distribution**) of the random variable.

Example 1 *Example random variables and distributions*

1.2 Distribution functions and probability (density) functions

To describe the distribution of a random variable X, it is sufficient to specify its **distribution function (d.f.)** (or **cumulative distribution function**), defined by

$$F_X(x) = P(X \leq x) \qquad \text{for all real } x.$$

Then other probabilities may be evaluated using the axioms of probability and the theorems which follow from them; for example if $x < y$ then

$$\begin{aligned} P(x < X \leq y) &= P(X \leq y) - P(X \leq x) \\ &= F_X(y) - F_X(x). \end{aligned}$$

A general distribution function F has the following properties.

1. $0 \leq F(x) \leq 1$ with $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1$.
2. $F(x)$ is non-decreasing in x; that is, if $x < y$ then $F(x) \leq F(y)$.

Those of you taking the Analysis module should note that a distribution function must also satisfy a property related to continuity:

3. F is **right-continuous** and has **left hand limits**. In other words, if $y \to x$ from above then $F(y) \to F(x)$, whereas if $y \to x$ from below then $F(y)$ approaches a limit which may or may not be equal to $F(x)$; it is denoted by $F(x-)$.

Most distributions which we will encounter are of two special types.

Discrete case

If X can only take integer values (or possibly values in some similar discrete set) then F increases entirely by jump discontinuities at these values and remains constant in between them; the size of the jump at value x will be

$$p(x) = P(X = x) = F(x) - F(x-)$$

and is called the **probability function (p.f.)** of X evaluated at x. In this case, probabilities may be found by summing the appropriate values of the probability function.

Absolutely continuous case

If F is continuous everywhere and differentiable (except possibly at a finite number of points) then its derivative

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

is called the **probability density function (p.d.f.)** of X. In this case probabilities may be found by integrating the p.d.f. over the appropriate range:

$$P(x < X \leq y) = F(y) - F(x) = \int_x^y f(t)\, dt.$$

Note that f must be non-negative because F is non-decreasing, but f is not itself a distribution function. For example, it is possible (and common) for f to be greater than 1 for some values of x.

Example 2 *Distribution functions and probability density functions*

1.3 Moments

The **mean** (or **expectation** or **expected value**) of a random variable X is defined as

$$\mu = \mu_X = E(X) = \begin{cases} \sum_{x \in R_X} xp(x) & \text{(discrete case);} \\ \int_{R_X} xf(x)\, dx & \text{(continuous case).} \end{cases}$$

Here R_X denotes the set of all values which X can take, known as the **range** of X.

More generally, if $g(X)$ is a function of X then

$$E\{g(X)\} = \begin{cases} \sum_{x \in R_X} g(x)p(x) & \text{(discrete case);} \\ \int_{R_X} g(x)f(x)\, dx & \text{(continuous case).} \end{cases}$$

Of particular interest is the **variance**

$$\sigma^2 = \sigma_X^2 = \text{Var}(X) = E(X-\mu)^2 = E(X^2) - \mu^2$$

and its positive square root σ, the **standard deviation**.

The mean is intended to be interpreted as a long-term average value of X. In fact the **weak law of large numbers** (MAS113, section 5.2) tells us that if we have a sequence of independent random variables $X_1, X_2, X_3, \ldots$ with the same distribution and with mean μ, and we take the average of the first n terms, $\bar{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}$, then for any $\epsilon > 0$, as $n \to \infty$

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0$$

and so for large n we expect $\bar{X}_n$ to be close to μ.

The mean and variance are special cases of **moments**; in general for any positive integer r we define the **rth moment of X about the origin** (or just the rth moment) as

$$\mu_r' = E(X^r)$$

and the r^{th} **moment of X about the mean** as

$$\mu_r = E(X-\mu)^r.$$

Thus $\mu = \mu_1'$ and $\sigma^2 = \mu_2$.

The third moment is used in defining the **coefficient of skewness** as

$$\beta_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{E(X-\mu)^3}{\sigma^3} = E\left(\frac{X-\mu}{\sigma}\right)^3.$$

This is a (dimensionless) quantity which tends to be positive if the distribution is positively skewed and negative if the distribution is negatively skewed; if it is symmetrical about μ then $X - \mu$ and $\mu - X$ have the same distribution and so

$$E(X-\mu)^3 = E(\mu - X)^3 = -E(X-\mu)^3,$$

so

$$E(X-\mu)^3 = 0.$$

So symmetry implies zero coefficient of skewness, as long as the third moment actually exists.

1.3.1 Random variables without a mean

The sum or integral in the definition of the mean may not converge; if it does not, we say that the mean does not exist (or sometimes that is infinite, but this can be misleading).

For example, let X be a random variable with probability density function

$$f(x) = \frac{1}{\pi(1+x^2)}.$$

This distribution is called the **Cauchy distribution** (and is also the special case of the Student t distribution, which you will have met in MAS113, with 1 degree of freedom). If we attempt to calculate the mean, we look at

$$\int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,dx,$$

which should be interpreted as

$$\lim_{s\to\infty, t\to\infty} \int_{-s}^{t} \frac{x}{\pi(1+x^2)}\,dx.$$

However

$$\int_{-s}^{t} \frac{x}{\pi(1+x^2)}\,dx = \frac{1}{2}\left(\log(1+t^2) - \log(1+s^2)\right),$$

and this does not have a well-defined limit as both s and t go to infinity. Hence the mean is undefined.

As mentioned above, the Weak Law of Large Numbers states that if we have a sequence of independent random variables $X_1, X_2, X_3, \ldots$ with the same distribution and with mean μ, then for any $\epsilon > 0$, as $n \to \infty$

$$P(|\bar{X}_n - \mu| > \epsilon) \to 0.$$

For a distribution without a defined mean, this result no longer makes sense, as we have no μ. If $X_1, X_2, X_3, \ldots, X_n$ are independent random variables with a Cauchy distribution, it turns out that $\bar{X}_n = \sum_{i=1}^{n} \frac{X_i}{n}$ in fact has a Cauchy distribution itself, regardless of the value of n, so there is no value that the sample mean is close to for large n. Similarly the Central Limit Theorem does not apply to random variables without a defined mean and variance.

The Cauchy distribution is not the only example of a distribution without a defined finite mean; a discrete example appears in the exercises.

1.4 Standard Distributions

There are many standard forms of distributions which occur in many different contexts; you will already have met some of them in MAS113. Each form is a family of distributions sharing a common formula for the p.f. or p.d.f. which contains one or more **parameter(s)**, the actual member of the family being determined by the value(s) of the parameter(s). (For example, the binomial family $Bi(n, p)$ has two parameters, n, the number of trials, and p, the success probability.)

Standard distributions are important either because they arise from simple models (e.g. the binomial distribution from Bernoulli trials) or because they have special mathematical properties (e.g. the normal distribution from the central limit theorem).

Two handouts will be made available, one with a list of standard distributions for discrete random variables, and another with a list of standard continuous distributions.

1.4.1 Standard discrete distributions

You will already have met some of the most important discrete distributions in first-year courses:

- The Bernoulli distribution
- The Binomial distribution
- The Poisson distribution
- The Geometric distribution

The following examples introduce a couple more standard discrete distributions, both of which are related to distributions you are already familiar with.

Example 3 *Hypergeometric distribution*

Example 4 *Negative binomial distribution*

All these distributions are described in the handout on standard discrete distributions.

1.4.2 The univariate normal distribution

Again, you will have encountered the normal distribution in MAS113.

If X has a normal distribution with mean μ and variance σ^2, we write $X \sim N(\mu, \sigma^2)$, and the probability density function of X is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

It can be shown by integration (see MAS113, section 4.6.2) that the mean and variance of a random variable with this p.d.f. really are μ and σ^2.

The special case $N(0, 1)$, with p.d.f.

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\},$$

is referred to as the **standard normal distribution**.

An important property of the normal distribution family is that if $X \sim N(\mu, \sigma^2)$ and a and b are constants, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. In particular X can be **standardised** by letting $Z = \frac{X-\mu}{\sigma}$, so that $Z \sim N(0, 1)$.

Another important property is that if we have n **independent** normal random variables $X_1, X_2, \ldots X_n$ with $X_i \sim N(\mu_i, \sigma_i^2)$ then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

(There will be some discussion of situations where the normal random variables are not independent later in this course.)

Example 5 *Finding $E(e^X)$ where X has a Normal distribution*

1.4.3 A note on the gamma and beta functions

The gamma and beta functions appear in the probability density functions of certain standard distributions and in this context can be thought of as normalising constants ensuring that the p.d.f.s integrate to 1. Both are defined as integrals.

The **gamma function** can be thought of as a generalisation of the factorial, and is defined by

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u}\, du$$

for $\alpha > 0$. Note that

$$\Gamma(1) = \int_0^\infty e^{-u}\, du = [-e^{-u}]_0^\infty = 1.$$

It is not hard to show using integration by parts that, for $\alpha > 1$,

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

and so if α is a positive integer then this may be iterated giving

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\Gamma(\alpha-2) = \ldots = (\alpha-1)!.$$

However, in general the integration cannot be performed explicitly. One other specific value, which appears in some formulae for standard distributions, is $\Gamma(1/2) = \sqrt{\pi}$.

Frequently in this course, we will encounter integrals of the form

$$\int_0^\infty u^{\alpha-1} e^{-\beta u}\, du$$

which are similar to the integral defining the Gamma function above, but which have an extra constant β. These can be related to the Gamma function by the following change of variables.

Lemma 1.4.1 *If $\beta > 0$, we have*

$$\int_0^\infty u^{\alpha-1} e^{-\beta u}\, du = \frac{\Gamma(\alpha)}{\beta^\alpha}.$$

Proof: We apply a change of variables $t = \beta u$; under this substitution

$$\begin{aligned} \int_0^\infty u^{\alpha-1} e^{-\beta u}\, du &= \int_0^\infty \left(\frac{t}{\beta}\right)^{\alpha-1} e^{-t} \left(\frac{1}{\beta}\right) dt \\ &= \beta^{-\alpha} \int_0^\infty t^{\alpha-1} e^{-t}\, dt \\ &= \beta^{-\alpha}\Gamma(\alpha). \end{aligned}$$

■

In a similar way the **beta function** is defined by

$$B(\alpha, \beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\, du$$

for $\alpha, \beta > 0$ and it can be shown (by a change of variables in a double integral) that it can be expressed in terms of the gamma function as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

1.4.4 The chi-squared distribution

First, consider the distribution of X^2 when $X \sim N(0,1)$. This is a skew distribution, the **chi-squared distribution with** 1 **degree of freedom**, with the following density

$$f(y) = \begin{cases} \frac{1}{\sqrt{2}\Gamma(1/2)} \frac{1}{\sqrt{y}} \exp\left(-\frac{y}{2}\right) & y > 0 \\ 0 & y < 0. \end{cases}$$

(We will show that this really is the correct density later in the course.)

In this case we write $Y = X^2 \sim \chi_1^2$ (the chi-square distribution with 1 degree of freedom). This distribution is a special case of a more general family of distributions, which you will have seen in MAS113, namely the chi-square distributions with n degrees of freedom, with p.d.f.

$$f(y) = \begin{cases} \frac{1}{\sqrt{2^n}\Gamma(n/2)} y^{n/2-1} \exp\left(-\frac{y}{2}\right) & y > 0 \\ 0 & y < 0. \end{cases}$$

This is the distribution of the sum of the squares of n independent standard normal random variables, i.e. if $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim N(0,1)$ and $Y = \sum_{i=1}^n X_i^2$, then $Y \sim \chi_n^2$. (It is possible to show this using methods from later in this course; see exercise 29.)

The chi-squared distribution is a special case of a more general family of distributions, the Gamma family.

1.4.5 The gamma distribution

Let

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

for $x \geq 0$ and $f(x) = 0$ for $x < 0$.

We first show that f is a probability density function. Obviously $f(x) \geq 0$. Now, using Lemma 1.4.1,

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\, dx &= \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{\beta^\alpha} \\ &= 1. \end{aligned}$$

So $f(x)$ is a p.d.f. The distribution with this p.d.f. is called the **Gamma distribution** with parameters α and β; if X has this p.d.f. we can write $X \sim Ga(\alpha, \beta)$. (NB there are alternative parametrisations of this distribution, so you sometimes have to be careful with software when using it.)

Example 6 *Find the mean and variance of X if $X \sim Ga(a, \beta)$.*

Notes:

- If $\alpha = 1$ we obtain the exponential distribution with parameter β.

- If $\alpha = \nu/2$ and $\beta = 1/2$ we obtain the chi squared distribution with ν degrees of freedom.
- If $X_1 \sim Ga(\alpha_1, \beta)$ and $X_2 \sim Ga(\alpha_2, \beta)$ and X_1 and X_2 are independent, then $X_1 + X_2 \sim Ga(\alpha_1 + \alpha_2, \beta)$. (See Example 19 later in the course.) It follows that a sum of independent exponential random variables with the same parameter has a Gamma distribution, and also that the sum of independent chi-squared random variables has a chi-squared distribution.

1.4.6 The beta distribution

Now we let

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 \leq x \leq 1$, and $f(x) = 0$ otherwise.

Again we start by showing that $f(x)$ is a p.d.f. Obviously $f(x) \geq 0$, and

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\, dx &= \int_0^1 \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}\, dx \\ &= 1, \end{aligned}$$

by the definition of the Beta function.

So $f(x)$ is a p.d.f. The distribution with this p.d.f. is called the **Beta distribution** with parameters α and β; if X has this p.d.f. we can write $X \sim Be(\alpha, \beta)$.

Example 7 *Calculate the mean and variance of X if $X \sim Be(\alpha, \beta)$*

Notes:

- The Beta distribution can be useful for modelling random quantities which are naturally constrained to be in $[0, 1]$ (or, via suitable scaling, in any fixed interval).
- The $Be(1, 1)$ distribution is the same as the Uniform distribution on $[0, 1]$.

1.5 Plotting distributions in R

The aim of this section is to show how to use the computer package R to plot density and distribution functions of random variables.

Most of you will have seen R in use in Level 1 courses. It is a statistical computing system which implements a dialect of the S language. For a more detailed introduction to R, including information on how to install it on your own computer, see the separate handout “An Introduction to R”.

To start with, we assume that R has been installed. Suppose we wish to plot the p.d.f. $f_X(x)$ of the random variable $X \sim N(0, 1)$. The command we use here is `curve`, which creates a curve of a given function. The form of this command is

```
> curve(f(x), from="lower limit", to="upper limit")
```

or just

```
> curve(f(x), "lower limit", "upper limit")
```

This tells R to plot a curve of a given function $y = f(x)$, where x takes values from "lower limit" to "upper limit". If we don't enter these two limits R will use its own default values. More details on the arguments of the above command can be found by typing `help(curve)`.

Using the command `curve` we can do

```
> curve(dnorm,-3,3)
```

Similarly, one can produce plots of the p.d.f. of any normal variable, $X \sim N(\mu, \sigma^2)$, by using the command `dnorm(x,mean,sd)`. For example

```
> curve(dnorm(x,2,10),-10,14)
```

gives a plot of the p.d.f. of a $N(2, 100)$ variable.

Similar plots can be obtained by finding out about the p.d.f.s of other distributions. It can be useful to use R's help system, which can be accessed with `help`(topic). For the normal distribution use `dnorm`, for the chi-square use `dchisq`, for the Student t use `dt`, for the gamma distribution use `dgamma` (but check the definition of the p.d.f., because there are different ways of paremetrising this distribution), for the beta distribution use `dbeta`, for the binomial use `dbinom`.

For example pictures of the p.d.f. of the chi-square $X \sim \chi_5$ and of the beta $Y \sim \text{Beta}(2, 3)$ distributions can be obtained by the following commands

```
> par(mfrow=c(1,2))
> curve(dchisq(x,5),0,40)
> curve(dbeta(x,2,3),0,1)
```

The result of these commands is shown in Figure 1.2. The first command tells R to print the two curves side by side.

For the above distributions, we can easily produce graphs of the distribution function by using `pnorm`, `pchisq`, etc. For example for the binomial, a graph of the distribution function can be obtained, as in Figure 1.3, by the command

```
> curve(pbinom(x,5,0.7),-1,8)
```

while the distribution function of the normal $N(0, 1)$ can be produced by the command

```
> curve(pnorm,-5,5)
```

and is shown in Figure 1.4.

If the distribution we wish to plot does not exist by default in R, then we can define it in R and plot it using the `curve` command as above. For more information on this, consult the on-line manuals of R.

1.6 Transformations of random variables

The general question here is: if we have a random variable X with a known distribution, and we have another random variable Y defined as

$$Y = g(X)$$

for some function $g : \mathbb{R} \to \mathbb{R}$, then what is the distribution of Y? For example, in §1.4.4 we stated that if X is standard normal and $Y = X^2$ then Y has χ_1^2 distribution. Another example is that if $X \sim N(\mu, \sigma^2)$, then $Y = (X - \mu)/\sigma \sim N(0, 1)$. In this section we will deal with transformations where both X and Y are continuous random variables.

If g is strictly **monotonic**, i.e. increasing or decreasing on the relevant range of x (which is not the case in the transformation from normal to chi-squared) then there is a method operating directly with p.d.f.'s which we derive below. Note that in this case g has an inverse function g^{-1} which is also increasing or decreasing as appropriate.

Firstly, in the increasing case, we may write

$$F_Y(y) = P(Y \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y)).$$

Differentiating,

$$f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y).$$

Similarly in the decreasing case we have

$$F_Y(y) = P(Y \le y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y)).$$

Hence

$$f_Y(y) = -f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y).$$

To cover both cases, we write

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| = f_X(g^{-1}(y)) \left| \frac{dx}{dy} \right|.$$

Example 8 *Transformation of a Gamma distribution*

If g is not monotonic then we need to be more careful. This applies (for example) to the transformation mentioned above: taking the square of a normal random variable to obtained a chi squared random variable.

Example 9 *Square of a standard normal*

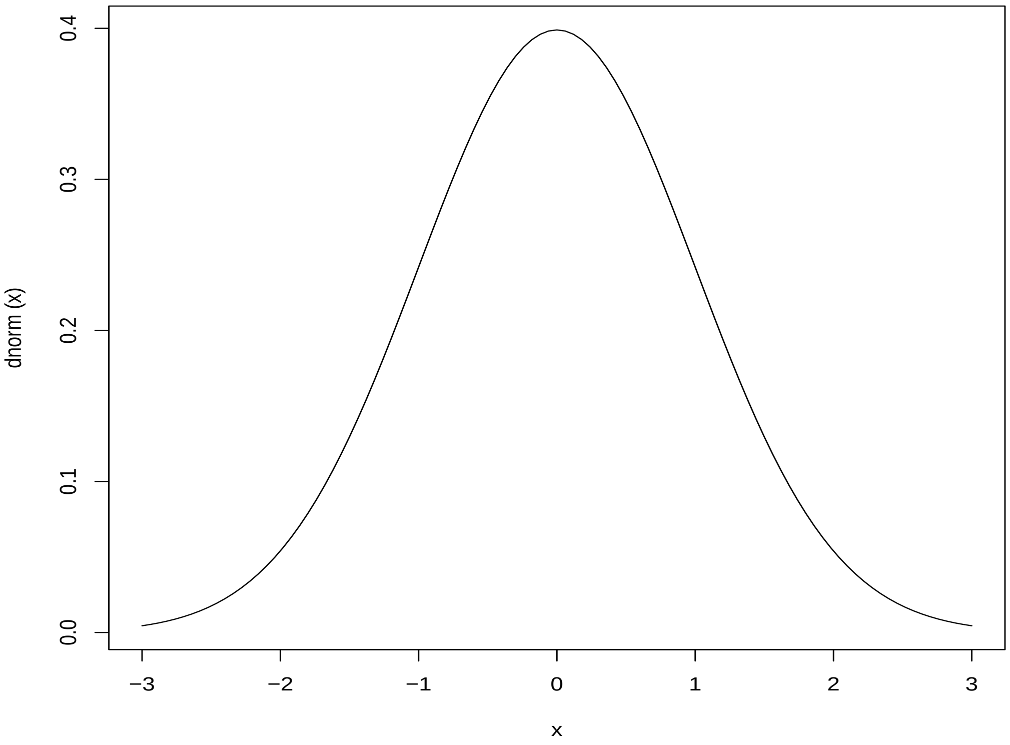

Figure 1.1: p.d.f. of a N(0,1) variable.

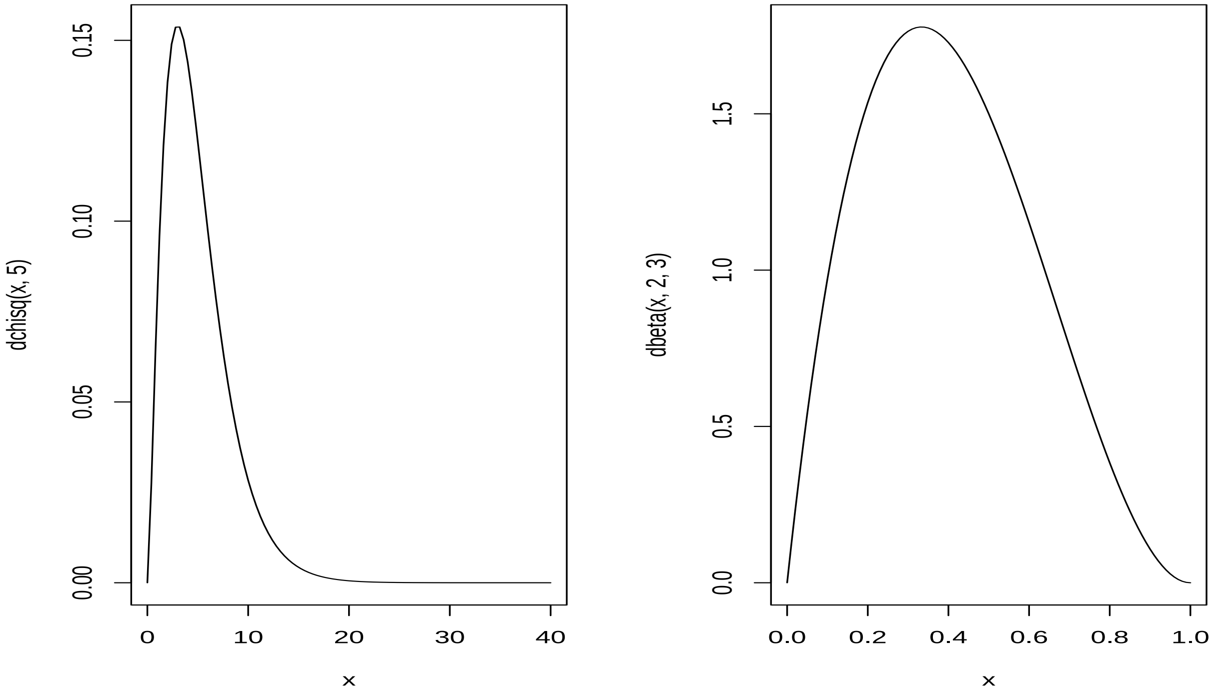

Figure 1.2: p.d.f. of a chi-square random variable (left panel) and of a beta variable (right panel).

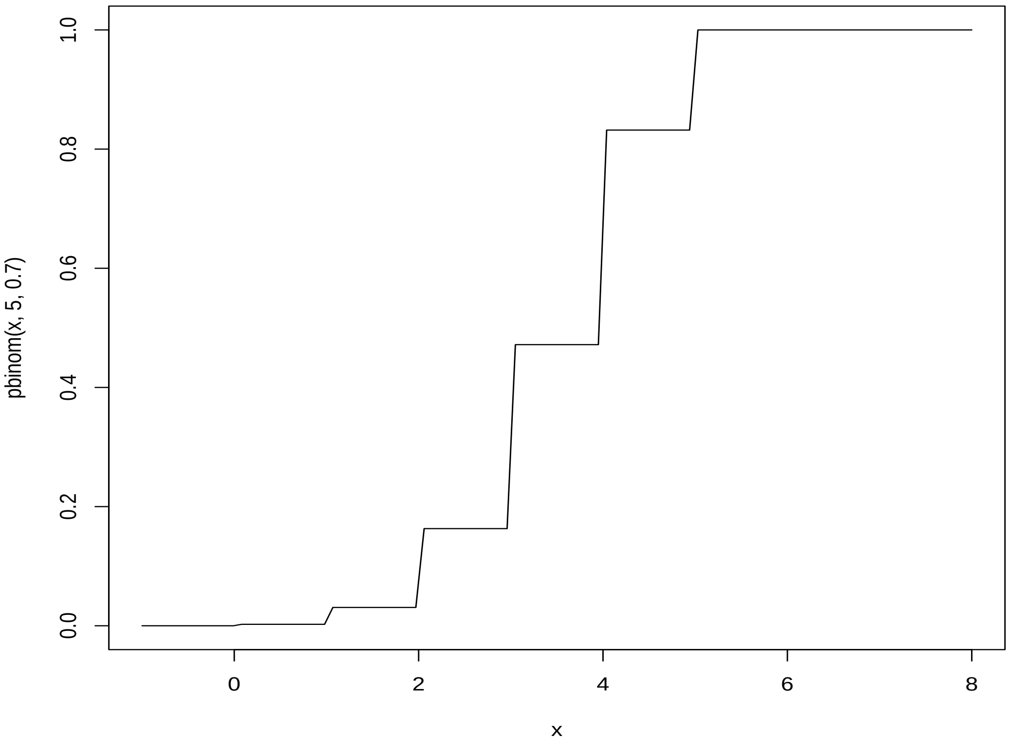

Figure 1.3: Distribution function of a binomial Bin$(5, 0.7)$ random variable.

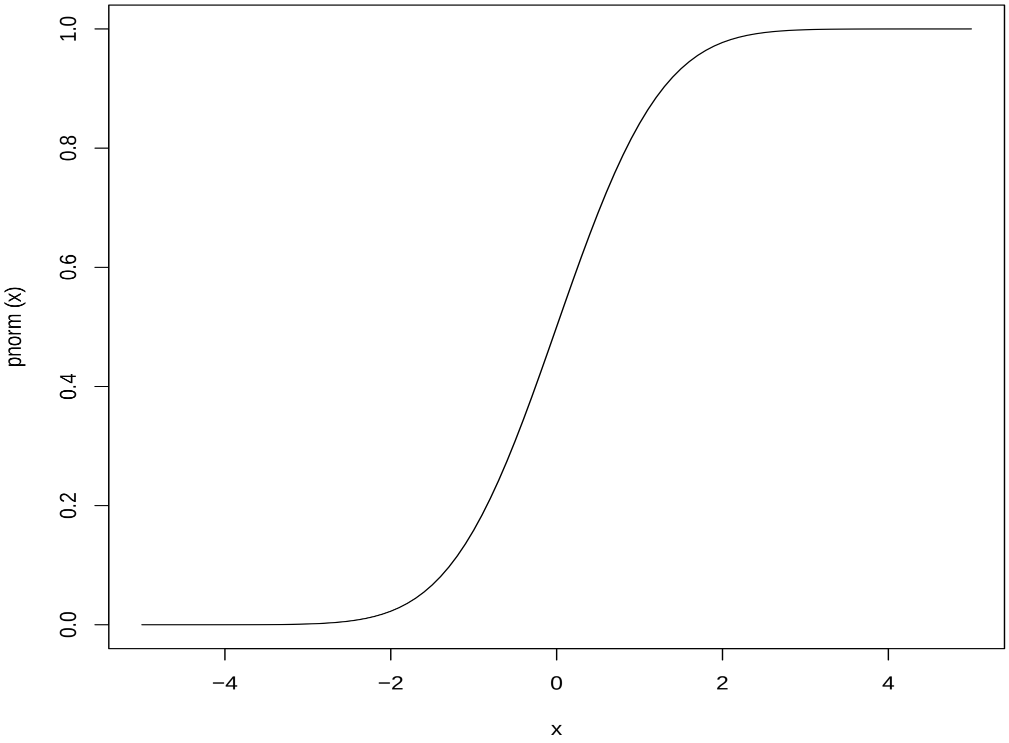

Figure 1.4: Distribution function of a normal $N(0,1)$ random variable.

Chapter 2

Multivariate distributions

2.1 Introduction

If we observe several (k) numerical quantities in the same experiment, for example for lung cancer patients

$$\begin{aligned} X_1 &= \text{age} \\ X_2 &= \text{size of tumour} \\ X_3 &= \text{smoking level} \\ X_4 &= \text{socio-economic group} \end{aligned}$$

then we have a **multivariate** random variable or random **vector**

$$\mathbf{X} = (X_1, X_2, \ldots, X_k)^T.$$

Formally, $\mathbf{X}$ is a mapping from the sample space S into k-dimensional space $\mathbb{R}^k$. Of particular interest will be how the components $X_1, X_2, \ldots, X_k$ vary together. If all $X_1, X_2 \ldots, X_k$ are continuous random variables $\mathbf{X}$ is said to be a continuous random vector; if all $X_1, \ldots, X_k$ are discrete random variables $\mathbf{X}$ is said to be a discrete random vector; in any other case it is said that $\mathbf{X}$ is neither continuous nor discrete.

You will already have met discrete random vectors in section 3.5 of MAS113 Introduction to Probability and Statistics; in this course we will extend the ideas there to discuss continuous random vectors, and we will introduce the important case of vectors with multivariate normal distributions. We will also look at transformations of multivariate distributions.

Often we will concentrate on the **bivariate** case $k = 2$ and denote the random variables by X and Y.

2.2 Continuous random vectors

2.2.1 Joint distribution and density functions

If X and Y are any two jointly distributed random variables then we can define their **joint distribution function** as

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

for all real x and y. In principle if we know the value of this function for all x, y then we can evaluate all other probabilities involving X and Y.

If $F_{X,Y}$ is sufficiently smooth to possess the partial derivative

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y) = f_{X,Y}(x, y)$$

(except possibly, for example, on the boundary of a region in which it is non-zero) then $f_{X,Y}$ is called the **joint probability density function (p.d.f.)** of X and Y. It is analogous to the p.d.f. of a univariate distribution, and may be thought of as measuring the "probability per unit area" at each point (x, y) in the plane.

To find the probability that the pair (X, Y) lies in some region D of the plane then we must integrate $f_{X,Y}$ over D; in other words

$$P((X, Y) \in D) = \int \int_D f_{X,Y}(x, y)\, dx\, dy. \tag{2.1}$$

Pictorially, if we plot the surface $z = f_{X,Y}(x, y)$ in three dimensions then this probability is the **volume** between this surface and the plane $z = 0$ determined by the set D in the (x, y) plane.

If $f_{X,Y}$ is a join probability density function of the random vector (X, Y) then, necessarily, we have

1. $f_{X,Y}(x, y) \geq 0$ for all x, y.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx\, dy = 1$.

Note that if we choose $D = (-\infty, x] \times (-\infty, y]$ in (2.1) we get

$$\begin{aligned} F_{X,Y}(x, y) &= P(X \leq x, Y \leq y) = P((X, Y) \in D) \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)\, du\, dv. \end{aligned}$$

Since evaluating probabilities involves double integration, often over a bounded region, it is important to get the **limits** of integration right; for example if we integrate with respect to y first then we need to ascertain the limits of y for each fixed x.

Example 10 *Joint probability density function; calculating probabilities*

2.2.2 Marginal distributions

Where we have a multivariate random variable $(X_1, X_2, \ldots, X_k)$, the **marginal distribution** of a component X_i is simply the distribution of X_i considered as a univariate random variable.

We can find the marginal distribution of the component of interest by "integrating out" the other variables. In the bivariate case (X, Y), the marginal p.d.f. of X is found by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy$$

("integrate y out"), remembering that if $f_{X,Y}(x, y)$ is positive on a restricted region then the effective limits of integration may depend on the value of x.

2.2.3 Conditional distributions

The conditional p.d.f. of Y given $X = x$ is given by the ratio of the joint p.d.f. and the marginal p.d.f. of the variable being conditioned on:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

provided $f_X(x) > 0$.

Example 11 *Marginal and conditional distributions*

2.2.4 Covariance and correlation

Let us write $\mu_X = E(X), \mu_Y = E(Y)$. The **covariance** $\mathrm{Cov}(X,Y)$ is defined as

$$\mathrm{Cov}(X,Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - E(X)E(Y)$$

where $E(XY)$ must be calculated as

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y) \; dx \; dy.$$

The **correlation coefficient** $\rho(X,Y)$ is then defined as

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{(\mathrm{Var}(X)\,\mathrm{Var}(Y))}}.$$

These measure the extent to which X and Y vary together.

(If you have a set of data which are a random sample from the distribution of (X,Y), the **Pearson's sample correlation coefficient** some of you may have seen is an estimate of $\rho(X,Y)$.)

Example 12 *Covariance and correlation*

2.2.5 Independence

If we have independent random variables X and Y with p.d.f.s $f_X(x)$ and $f_Y(y)$ respectively, then the random vector (X,Y) has joint p.d.f. given by $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Furthermore, if we know $f_{X,Y}$ we can determine whether X and Y are independent; in fact X and Y are independent if and only if there exist a pair of functions $g(x)$ and $h(y)$ such that $f_{X,Y}$ factorises into the form

$$f_{X,Y}(x,y) = g(x)h(y)$$

for all x, y. In this case g and h are necessarily the marginal p.d.f.'s of X and Y, because

$$f_X(x) = \int_{-\infty}^{\infty} g(x)h(y) \; dy = g(x) \int_{-\infty}^{\infty} h(y) \; dy = g(x).$$

Example 13 *Independence*

2.3 Conditional expectation

The **conditional expectation** of Y given that X takes the value x can be defined as the expectation of a random variable with the conditional distribution:

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\, dy.$$

Note that the formula above is a function of x; we could write it as $g(x) = E(Y|X=x)$. We define

$$\mathbb{E}(Y|X) = g(X),$$

which formally means that $E(Y|X)$ is a random variable and, for each element s of the sample space, $E(Y|X)(s) = g(X(s)) = E(Y|X = X(s))$.

Example 14 *Calculating conditional expectation*

We define the conditional variance as the variance of the conditional distribution of Y given X. In symbols,

$$\text{Var}(Y|X) = E\{(Y - E(Y|X))^2|X\}.$$

We can also define conditional covariances: if X, Y and Z are random variables, then the conditional covariance of X and Y, given Z, is

$$\text{Cov}(X, Y|Z) = E(XY|Z) - E(X|Z)E(Y|Z).$$

2.3.1 Properties of conditional expectations

1. $E(Y) = E\{E(Y|X)\}$.
2. $\text{Var}(Y) = E\{\text{Var}(Y|X)\} + \text{Var}\{E(Y|X)\}$.
3. $\text{Cov}(X, Y) = E\{\text{Cov}(X, Y|Z)\} + \text{Cov}\{E(X|Z), E(Y|Z)\}$.
4. $E(Yf(X)|X) = f(X)E(Y|X)$

Properties 1 and 2 are useful when the best way of finding the mean and variance of Y is by conditioning on X. Property 4 says that any factor which is a function of X only may be taken outside the expectation.

Example 15 *Proof of Property 1*

Example 16 *Calculation of expectation and variance by conditioning*

2.4 Transformations of multivariate distributions

We will concentrate here on the bivariate case, but the theory described extends to the more general case.

We are interested in the situation where we have two jointly distributed continuous random variables, say X and Y with joint p.d.f. $f_{X,Y}(x, y)$, and transforming them into two new continuous random variables, say U and V with joint p.d.f. $f_{UV}(u, v)$, given by say $U = g(X, Y)$ and

$V = h(X, Y)$, in such a way that the whole transformation is continuous, differentiable and **one-to-one** in that there exist "inverse" functions G and H with $X = G(U, V)$ and $Y = H(U, V)$.

If we take a small region around (x, y) then this is transformed into a small region around (u, v), where $u = g(x, y)$ and $v = h(x, y)$, and the area of this new region will be the area of the old region multiplied by the **Jacobian** of the transformation

$$\left| \det \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} \right|$$

evaluated at (x, y). Bearing in mind that a probability density function measures probability per unit area, in order to evaluate the joint p.d.f. of U and V at (u, v) we need to take the joint p.d.f. of X and Y at (x, y) and **divide** it by this Jacobian. In fact, since the joint p.d.f. is to be expressed in terms of u and v, it is (in most cases) easier to multiply by the Jacobian of the inverse transformation $x = G(u, v), y = H(u, v)$ where G and H are as defined above. So we get

$$f_{UV}(u, v) = f_{X,Y}(G(u, v), H(u, v)) \left| \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \right|.$$

This formula generalises the one obtained in the univariate case for monotonic g in Chapter 1.

It is important to identify the range of values taken by (U, V), possibly with the aid of a graph. In particular, if X and Y take values in a restricted range given by inequalities in x and y, then these must be translated into inequalities in u and v by substituting for x and y in terms of u and v.

Example 17 *Transforming bivariate random variables*

Example 18 *Application to simulation of normal random variables*

Sometimes we are interested in only one transformed random variable, $U = g(X, Y)$ say. In this case one possibility is to choose V arbitrarily (but not identical to or functionally dependent on U, to ensure that the joint distribution of U and V is genuinely two-dimensional), to find the joint p.d.f. of U and V, and then to eliminate the unwanted V by finding the marginal p.d.f. of U. If there is no other obvious choice, choosing $V = X$ or $V = Y$ often works well.

Example 19 *Finding the distribution of a sum of Gamma random variables*

Note that the method introduced in this section is closely related to the method used when changing variables in multiple integration.

2.4.1 An application: the Student t distribution

The **Student t distribution** (named after William Gosset, who used the pen-name "Student") arises when we have independent random variables $Z \sim N(0, 1)$ and $W \sim \chi^2_n$, and we consider the random variable

$$X = \frac{Z}{\sqrt{W/n}}.$$

We write $X \sim t_n$. As with the chi squared distribution, the parameter n is referred to as the number of degrees of freedom.

To find the probability density function of X, consider (Z, W) as a bivariate random vector, and transform it to (X, Y) where $X = \frac{Z}{\sqrt{W/n}}$ as above and $Y = W$. Then, by independence and the formulae for the p.d.f.s of standard Normal and chi squared distributions,

$$f_{Z,W}(z, w) = \frac{1}{\sqrt{2^n}\Gamma(n/2)\sqrt{2\pi}} w^{\frac{n}{2}-1} \exp\left(-\frac{(z^2 + w)}{2}\right),$$

for $w > 0$. If $x = \frac{z}{\sqrt{\frac{w}{n}}}$ and $y = w$ then the inverse transformation is given by $w = y$ and $z = x\sqrt{y/n}$, with Jacobian $\sqrt{y/n}$, so the joint p.d.f. of X and Y is

$$\begin{aligned} f_{X,Y}(x, y) &= \frac{1}{\sqrt{2^n}\Gamma(n/2)\sqrt{2\pi}} y^{\frac{n}{2}-1} \exp\left(-\frac{\left(\frac{x^2 y}{n} + y\right)}{2}\right) \sqrt{y/n} \\ &= \frac{1}{\sqrt{2^n}\Gamma(n/2)\sqrt{2\pi n}} y^{\frac{n-1}{2}} \exp\left(-\frac{y}{2}\left(\frac{x^2}{n} + 1\right)\right), \end{aligned}$$

for $y > 0$.

To obtain the p.d.f. of X, integrate out y:

$$\begin{aligned} f_X(x) &= \int_0^\infty f_{X,Y}(x, y)\, dy \\ &= \frac{1}{\sqrt{2^n}\Gamma(n/2)\sqrt{2\pi n}} \int_0^\infty y^{\frac{n-1}{2}} \exp\left(-\frac{y}{2}\left(\frac{x^2}{n} + 1\right)\right)\, dy. \end{aligned}$$

The integral here is, by Lemma 1 on the Gamma function (section 1.4.3),

$$\Gamma((n+1)/2)\left(\frac{1}{2}\left(\frac{x^2}{n} + 1\right)\right)^{-(n+1)/2},$$

giving

$$f_X(x) = \frac{1}{\sqrt{2^n}\Gamma(n/2)\sqrt{2\pi n}} \Gamma((n+1)/2)\left(\frac{1}{2}\left(\frac{x^2}{n} + 1\right)\right)^{-(n+1)/2},$$

which simplifies to

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}\left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \qquad (2.2)$$

From this formula, it's possible to see that if $n \geq 2$ the mean is $E(X) = 0$ and, if $n \geq 3$ the variance is $\text{Var}(X) = \frac{n}{n-2}$. Additionally, as $n \to \infty$ $f(x)$ converges to the p.d.f. of a standard normal distribution.

You will have seen this distribution before, in MAS113 (sections 6 and 7) where the t-test was introduced. Describing the Student t distribution as $Z/\sqrt{W/n}$ makes it seem somewhat unnatural, at first glance. However, the Student t distribution appears naturally in statistics, for the following reason.

Let $Z_1, Z_2, \ldots$ be independent identically distributed normal random variables with mean μ and variance σ^2. Let us now regard n samples of these as a set of data. The sample mean and variance are respectively $\bar{Z} = \frac{1}{n}(Z_1 + \ldots, Z_n)$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \mu)^2$. Then, it turns out that the statistic

$$X' = \frac{\bar{Z} - \mu}{s/\sqrt{n}}$$

has the Student t distribution with $n-1$ degrees of freedom (it takes a bit of effort to prove this, see question 37!). Crucially, as can be seen from (2.2), the distribution of X' does not depend on μ or σ. This unusual property makes X' a very useful statistic.

2.5 Covariance matrices and linear transformations

Let $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ be a random (column) vector with **mean**

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)^T = (E(X_1), E(X_2), \ldots, E(X_k))^T = E(\mathbf{X}).$$

Then the $k \times k$ matrix Σ with elements given by

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = E((X_i - \mu_i)(X_j - \mu_j))$$

for $i, j = 1, 2, \ldots, k$ is called the **covariance matrix** of $\mathbf{X}$, denoted by $\mathrm{Cov}(\mathbf{X})$. This matrix has the variances $\sigma_1^2, \sigma_2^2, \ldots \sigma_k^2$ of the random variables down the diagonal, and is **symmetric**, because $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$. We may also write

$$\mathrm{Cov}(\mathbf{X}) = E\left((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right)$$

where the expectation is taken componentwise. Note that here, $(\mathbf{X} - \boldsymbol{\mu})$ is a $1 \times k$ matrix and $(\mathbf{X} - \boldsymbol{\mu})^T$ is a $k \times 1$ matrix; they are multiplied together as matrices to give a $k \times k$ matrix.

From the definition of correlation coefficient we may also write $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ where ρ_{ij} is the correlation coefficient between X_i and X_j.

If $X_1, X_2, \ldots X_k$ are **independent** (or merely uncorrelated) then Σ is a **diagonal** matrix (having zero off-diagonal elements).

Example 20 *Example of a covariance matrix*

Matrix notation is useful when we consider linear transformations of $\mathbf{X}$. Let A be a fixed $m \times k$ matrix and $\mathbf{b}$ be a fixed m-vector, and write

$$\mathbf{Y} = A\mathbf{X} + \mathbf{b}$$

so that $\mathbf{Y}$ has m components. Then since pre-multiplying by a matrix is a linear operation we get

$$\begin{aligned} E(\mathbf{Y}) &= AE(\mathbf{X}) + \mathbf{b} \\ E(\mathbf{Y}) &= A\boldsymbol{\mu} + \mathbf{b}. \end{aligned}$$

Also

$$\begin{aligned} \mathrm{Cov}(\mathbf{Y}) &= E\left((\mathbf{Y} - A\boldsymbol{\mu} - \mathbf{b})(\mathbf{Y} - A\boldsymbol{\mu} - \mathbf{b})^T\right) \\ &= E\left(A(\mathbf{X} - \boldsymbol{\mu})(A(\mathbf{X} - \boldsymbol{\mu}))^T\right) \\ &= E\left(A(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T A^T\right) \\ &= AE\left((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right) A^T \\ \mathrm{Cov}(\mathbf{Y}) &= A\Sigma A^T. \end{aligned}$$

Example 21 *Linear transformation of a random vector*

Variance of linear combinations

We can use the above theory to obtain a formula for the variance of a linear combination of the random variables $X_1, X_2 \ldots, X_k$, perhaps with a constant term, say $Y = a_1X_1 + a_2X_2 + \ldots + a_kX_k + b$. We do this by choosing $m = 1$ and letting A be a row vector with appropriate entries, $\mathbf{a}^T = (a_1, a_2, \ldots, a_k)$, so that $A\mathbf{X} + b$ is the scalar $Y = a_1X_1 + a_2X_2 + \ldots + a_kX_k + b$.

Using the previous theory, we get

$$\mathrm{Var}(Y) = \mathbf{a}^T\Sigma\mathbf{a}.$$

Since this is always non-negative, we have shown that Σ is a **positive semi-definite** matrix, i.e. one for which $\mathbf{a}^T\Sigma\mathbf{a} \geq 0$ for all $\mathbf{a}$. (A **positive definite** matrix is one where the inequality is strict for all non-zero $\mathbf{a}$.) Note that a positive semi-definite matrix has all its eigenvalues non-negative (which can be seen by letting $\mathbf{a}$ be an eigenvector in the definition).

A particular special case is the general formula for variance of a sum $X_1 + X_2 + \ldots + X_k$, covering cases where the variables in the sum are not necessarily independent. To do this, let each element of $\mathbf{a}$ be 1 and let $b = 0$. Then

$$\begin{aligned} \mathrm{Var}\sum_{i=1}^{k} X_i &= (1, 1, \ldots, 1)\Sigma(1, 1, \ldots, 1)^T \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_{ij} \\ &= \sum_{i=1}^{k} \mathrm{Var}(X_i) + 2 \sum_{i,j:1\leq i<j\leq k} \mathrm{Cov}(X_i, X_j). \end{aligned}$$

Example 22 *Variance of a sum*

2.6 The multivariate normal (MVN) distribution

This is the most important continuous joint distribution, and often is a natural choice for modelling multivariate data. For example, we might have data on irises, with measurements of sepal length, sepal width, petal length, petal width. For each of these individually, we would probably choose a (univariate) normal distribution as our model, and the multivariate normal distribution provides a way of modelling the way that they vary together where each of the marginal distributions is univariate normal.

2.6.1 The independent bivariate case

Consider two independent random variables U and V each following respectively (univariate) normal distributions, so that $U \sim N(\mu_1, \sigma_1^2)$ and $V \sim N(\mu_2, \sigma_2^2)$ with p.d.f.'s

$$f_U(u) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(u-\mu_1)^2}{2\sigma_1^2}\right\}$$

and

$$f_V(v) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(v-\mu_2)^2}{2\sigma_2^2}\right\}$$

By independence, the joint p.d.f. of U and V is given by the product of the individual p.d.f.s, so is

$$f_{U,V}(u,v) = f_U(u)f_V(v) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\left\{-\frac{1}{2}\left[\frac{(u-\mu_1)^2}{\sigma_1^2}+\frac{(v-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

for $(u,v)^T \in \mathbb{R}^2$. This is a first example of a multivariate normal.

2.6.2 The general bivariate case

Let the random vector $(U,V)^T$ be defined as in the previous section, with both means zero and both variances 1, so that

$$f_{U,V}(u,v) = \frac{1}{2\pi}\exp\left\{-\frac{1}{2}\left[u^2+v^2\right]\right\}.$$

Now let

$$S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$$

be a non-singular 2×2 matrix, and let $\boldsymbol{\mu} = (\mu_1,\mu_2)^T$ be a 2-vector. We now consider the random vector

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = S\begin{pmatrix} U \\ V \end{pmatrix} + \boldsymbol{\mu}.$$

We can consider this as a transformation of the random vector $\begin{pmatrix} U \\ V \end{pmatrix}$, and so we can use the theory given in section 2.4. The forward transformation is given by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = S\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

and the inverse transformation is given by

$$\begin{pmatrix} u \\ v \end{pmatrix} = S^{-1}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}.$$

Using the form of the inverse matrix, we can re-write the inverse transformation as

$$u = \frac{1}{\det S}(s_{22}(x_1-\mu_1)-s_{12}(x_2-\mu_2)), \quad v = \frac{1}{\det S}(-s_{21}(x_1-\mu_1)+s_{11}(x_2-\mu_2)),$$

and the Jacobian of the inverse transformation is $|1/\det S|$. Hence the joint p.d.f. of X_1 and X_2, $f_{X_1,X_2}(x_1,x_2)$, is

$$\frac{1}{2\pi|\det S|}\exp\left\{-\frac{\left[(s_{22}(x_1-\mu_1)-s_{12}(x_2-\mu_2))^2+(-s_{21}(x_1-\mu_1)+s_{11}(x_2-\mu_2))^2\right]}{2(\det S)^2}\right\},$$

which can be rearranged as

$$\frac{1}{2\pi|\det S|}\exp\left\{-\frac{\left[\sigma_2^2(x_1-\mu_1)^2+\sigma_1^2(x_2-\mu_2)^2-2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2)\right]}{2(\det S)^2}\right\},$$

where $\sigma_1^2 = s_{11}^2+s_{12}^2$, $\sigma_2^2 = s_{21}^2+s_{22}^2$ and $\sigma_{12} = s_{22}s_{12}+s_{21}s_{11}$.

The transformation we have used is linear, so we can also use the theory of section 2.5 to work out the mean vector and covariance matrix of $(X_1, X_2)^T$. They are

$$E\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = S\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$\begin{aligned} \Sigma = \operatorname{Cov}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} &= S\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}S^T = SS^T \\ &= \begin{pmatrix} s_{11}^2 + s_{12}^2 & s_{22}s_{12} + s_{21}s_{11} \\ s_{22}s_{12} + s_{21}s_{11} & s_{21}^2 + s_{22}^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}, \end{aligned}$$

where $\sigma_1^2, \sigma_2^2, \sigma_{12}$ are defined as above. (So σ_1^2 and σ_2^2 really are the variances of X_1 and X_2, and σ_{12} really is their covariance, as suggested by the choice of notation.)

Finally note that $\det\Sigma = \det(SS^T) = (\det S)^2$, so we can replace $|\det S|$ by $\sqrt{\det\Sigma}$ in the above. As $\det\Sigma = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$, we can re-write the joint p.d.f. f_{X_1,X_2} as

$$\frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}} \exp\left\{ -\frac{\sigma_2^2(x_1-\mu_1)^2 - 2\sigma_{12}(x_1-\mu_1)(x_2-\mu_2) + \sigma_1^2(x_2-\mu_2)^2}{2(\sigma_1^2\sigma_2^2 - \sigma_{12}^2)} \right\}, \qquad (2.3)$$

for all $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$.

By analogy with the univariate case we write $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

the covariance matrix, and we say that the random vector $\mathbf{X}$ follows the bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

Note that the joint p.d.f. can also be written in terms of the matrix Σ as

$$f(X_1, X_2) = \frac{1}{2\pi(\det(\Sigma))^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}.$$

This p.d.f. is well defined for any symmetric positive definite 2×2 matrix Σ.

Remark 2.6.1 *To see this, note that any such matrix (because it is symmetric) can be diagonalised as $\Sigma = PDP^{-1} = PDP^T$, where P is orthogonal, and D will have positive entries (because Σ is positive definite). So D can be written as $\hat{D}^2$, and $\Sigma = P\hat{D}\hat{D}P^T$. If we let $S = P\hat{D}Q$ for any orthogonal matrix Q, then (using orthogonality)*

$$SS^T = P\hat{D}Q(P\hat{D}Q)^T = P\hat{D}QQ^T\hat{D}P^T = P\hat{D}\hat{D}P^T = \Sigma.$$

So any positive definite Σ can be obtained by using a suitable S.

For a 2×2 symmetric matrix, being positive definite is equivalent to $\sigma_1^2, \sigma_2^2 > 0$ and $\sigma_{12}^2 < \sigma_1^2\sigma_2^2$. If Σ is a diagonal matrix, so $\sigma_{12} = 0$, then we recover the independent case.

The contours of f are concentric ellipses centred on $\boldsymbol{\mu}$.

2.6.3 Marginal distributions are normal

Taking marginal distributions preserves normality. To see this, using the derivation of the bivariate normal in section 2.6.2 we can write the components of a bivariate normal random vector $\mathbf{X}$ as $X_1 = s_{11}U + s_{12}V + \mu_1$ and $X_2 = s_{21}U + s_{22}V + \mu_2$ where U and V are independent standard normal random variables. The theory of the univariate normal distribution now tells us that the marginal distributions of the components are univariate normals: $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$.

It is also possible to use the method of section 2.2.2 directly with the bivariate normal p.d.f.; see exercise 33(a).

2.6.4 Correlation, covariance, and independence

If the components of a bivariate normal have covariance (and thus correlation) zero, then they are independent. This can be seen by letting $\sigma_{12} = 0$ in the form for the p.d.f. of a multivariate normal in (2.3); the joint p.d.f. then factorises into two univariate normal p.d.f.s.

It is important to remember that this result does not hold for general random variables, but it does hold for components of multivariate normals. It is possible to find pairs of random variables which are not independent but have correlation zero; see for example Exercise 24.

2.6.5 Linear transformations of the bivariate normal

Suppose $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$, for some known mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

For any non-singular 2×2 matrix A and for any 2×1 vector $\mathbf{b}$ define the linear transformation $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$.

From section 2.6.2, we know that $\mathbf{X} = S\mathbf{U} + \boldsymbol{\mu}$ for some matrix S and vector $\boldsymbol{\mu}$. So we can write

$$\mathbf{Y} = AS\mathbf{U} + A\boldsymbol{\mu} + \mathbf{b},$$

telling us that $\mathbf{Y}$ itself has a bivariate normal distribution with mean vector

$$\boldsymbol{\mu}_Y = E(\mathbf{Y}) = A\boldsymbol{\mu} + \mathbf{b} = AE(\mathbf{X}) + \mathbf{b}$$

and covariance matrix

$$AS(AS)^T = ASS^TA^T = A\,\mathrm{Cov}(\mathbf{X})A^T.$$

We can replace A here by a row vector, $\mathbf{b} = b$ by a scalar, so giving $Y = A\mathbf{X} + b$ as a scalar. Normality is again preserved, and the mean and the variance of Y are

$$\mu_Y = E(Y) = AE(\mathbf{X}) + b, \quad \sigma_Y^2 = \mathrm{Var}(Y) = A\,\mathrm{Cov}(\mathbf{X})A^T$$

giving a univariate normal distribution

$$\mathbf{Y} \sim N(\mu_Y, \sigma_Y^2).$$

Example 23 *Transformations of the bivariate normal*

2.6.6 Conditional distributions are normal

Taking conditional distributions preserves normality, so that if $\mathbf{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ then the conditional distribution of X_2 given $X_1 = x_1$ is a univariate normal distribution. To see this, let $Y_1 = X_1$ and $Y_2 = X_2 - \lambda X_1$ for some λ which we will choose later. Then

$$E(\mathbf{Y}) = \begin{pmatrix} \mu_1 \\ \mu_2 - \lambda\mu_1 \end{pmatrix}$$

and

$$\begin{aligned} \mathrm{Cov}(\mathbf{Y}) &= \begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 & -\lambda \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ -\lambda & 1 \end{pmatrix} \begin{pmatrix} \sigma_1^2 & -\lambda\sigma_1^2 + \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & -\lambda\rho\sigma_1\sigma_2 + \sigma_2^2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & -\lambda\sigma_1^2 + \rho\sigma_1\sigma_2 \\ -\lambda\sigma_1^2 + \rho\sigma_1\sigma_2 & \lambda^2\sigma_1^2 - 2\lambda\rho\sigma_1\sigma_2 + \sigma_2^2 \end{pmatrix}. \end{aligned}$$

Hence if we choose $\lambda = \rho\sigma_1^{-1}\sigma_2$ then the above covariance matrix simplifies to

$$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & (1-\rho^2)\sigma_2^2 \end{pmatrix}$$

and, in particular, Y_1 and Y_2 are independent. Hence we may write

$$X_2 = \lambda X_1 + Y_2$$

where Y_2 is independent of X_1. In particular, conditional on $X_1 = x_1$,

$$X_2 = \lambda x_1 + Y_2$$

and so X_2 is normally distributed with mean

$$\lambda x_1 + (\mu_2 - \lambda\mu_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

and variance $(1-\rho^2)\sigma_2^2$.

A particular consequence of the above result is that

$$E(X_2|X_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X_1 - \mu_1)$$

and

$$\mathrm{Var}(X_2|X_1) = (1-\rho^2)\sigma_2^2.$$

The conditional expectation depends linearly on X_1 and the conditional variance does not depend upon X_1.

See also Exercise 33(b) for an alternative method.

Example 24 *Conditional distributions for bivariate normal*

2.6.7 Higher dimensions

We now generalise to higher dimensions. Given a vector, $\boldsymbol{\mu}$, of length k and a $k \times k$ positive definite symmetric matrix Σ we can define the function

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}(\det(\Sigma))^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

It can be shown that this does indeed define a joint p.d.f. for any choice of $\boldsymbol{\mu}$ and Σ; this can be derived from the independent case by using a suitable transformation much as in section 2.6.2 for the bivariate case. The joint distribution so specified is called the **multivariate normal distribution** $N(\boldsymbol{\mu}, \Sigma)$ or $N_k(\boldsymbol{\mu}, \Sigma)$; it may be shown that it does indeed have mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

2.6.8 Transformations of the multivariate normal

Let $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \Sigma)$ and consider a transformation of the form $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ where A is a $m \times k$ matrix with $m \leq k$ and A is of full rank m so that $\mathbf{Y}$ has non-singular covariance matrix.

If $m = k$, so that A is a square matrix, then essentially the same argument as in section 2.6.5 for the bivariate case shows that

$$\mathbf{Y} \sim N_k(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T).$$

This preservation of normality extends to the case where $m < k$, where

$$\mathbf{Y} \sim N_m(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T).$$

In particular this property implies that all **marginal** distributions are (multivariate) normal; for example if $k = 5$ and we take $m = 2, \mathbf{b} = 0$ and

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

then we see that the marginal joint distribution of X_1 and X_2 is multivariate normal. As in the bivariate case, it is also possible to show this by integrating out variables.

2.6.9 Further properties of the multivariate normal

Conditional distributions preserve multivariate normality: the conditional distribution of a set of the components, given values for the remaining components, will have a multivariate normal distribution.

Components are independent if and only if their covariance is zero; again note that this is a special property of the multivariate normal.

Recall that the (univariate) normal p.d.f. cannot be integrated explicitly and that normal probabilities have to be approximated numerically and tabulated or evaluated using a computer package. Multivariate normal probabilities are even more difficult to evaluate unless the region of interest takes a special shape.

Example 25 *Example of a multivariate normal*

Chapter 3

Likelihood: Theory

3.1 The setting

In this section we will start to look at the idea of **statistical inference**, meaning methods of using analysis of data to obtain information about the processes which produced the data. In particular, we will be looking at methods of inference based on **likelihood**. Many common methods of data analysis rely on likelihood.

3.1.1 Data

Typically we will have a set of n data values which we can think of as a vector, $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. We will think of the data as being realisations of a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Note the use of capital letters for the random variables and lower case letters for the values they take.

The random vector $\mathbf{X}$ will have a joint p.d.f.

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \ldots, x_n).$$

Often it is clear which random variable the p.d.f. refers to, in which case we will drop the subscripts $\mathbf{X}$ and $X_1, X_2, \cdots, X_n$. This p.d.f. will be unknown, the aim of the inference being to obtain information about it.

We assume that our data $x_1, x_2, \ldots, x_n$ comes from independent, identically distributed experiments. Because of this, we also assume that the random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed. In this case, the joint p.d.f. $f_{\mathbf{X}}(\mathbf{x})$ will be a product of terms for each experiment:

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f(x_i), \tag{3.1}$$

where $f(x)$ is the common p.d.f. of the random variables $X_1, X_2, \ldots, X_n$.

If we have discrete random variables, then we would have a probability function instead of a p.d.f. The theory in this case is very similar, so it is common to use the same notation in both cases. We will do so, only in this chapter! Therefore, in discrete examples we will have $f_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$.

3.1.2 Models and Parameters

We assume that we already know that the joint p.d.f. of $\mathbf{X}$ takes a particular form, usually involving some standard distribution. We refer to this as our **model**. However, the parameters of this standard distribution are unknown, and our aim in analysing the data will be to obtain good choices of values for these parameters, based on the data we have[1].

For example, suppose we have a biased coin X, which shows either H or T. We don't know $\theta = \mathbb{P}[H]$, and would like to estimate it. We could flip the coin n times, generating data $x_1, x_2, \ldots, x_n$; we know that the number of heads follows a $\text{Bin}(n, \theta)$, distribution - but we don't know the value of θ. We can estimate θ from our data, using the methods in this chapter.

We denote the parameters of our model by $\boldsymbol{\theta}$; we represent $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$ as a vector, although in any particular case the set of parameters can be a scalar (a single number), a matrix or some other structure. We write Θ for the set of possible parameter values.

Sometimes some of the unknown parameters are so-called **nuisance parameters**: their values are unknown, and we have to take account of this in our analysis, but they are not what we are really interested in.

Given a model, and a particular set of parameter values $\boldsymbol{\theta}$, we write the p.d.f. (if we are dealing with continuous random variables) for $\mathbf{X}$ as $f(\mathbf{x}; \boldsymbol{\theta})$. (Again, we use the same notation in the discrete case, but it now means a probability function.)

Example 26 *Chemical reaction*

3.2 Likelihood

3.2.1 The likelihood function

If we have a family of distributions, parametrized by θ, with p.d.f.s (or p.f.s) $x \mapsto f(x; \theta)$ then the **likelihood** of θ given the observed outcome x is defined to be

$$L(\theta; x) = f(x; \theta).$$

The point here is that, for a fixed piece of data x, we regard $L(\theta; x)$ as a function of θ. We call $\theta \mapsto L(\theta; x)$ the **likelihood function**. Note that, considered this way, it is no longer a p.d.f.; in particular there is no requirement for it to integrate (over θ) to 1.

Example 27 *Binomial likelihood*

If we have a statistical model, $f_{\mathbf{X}}(\mathbf{x}; \theta)$, as in Section 3.1, with parameter values $\boldsymbol{\theta}$ then $\{f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\,;\, \boldsymbol{\theta} \in \Theta\}$ is the family of p.d.f.s (in the continuous case) that define our model.

Once we have observed a particular set of data values $\mathbf{x}$ we can consider $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. Since we assume that our data points are i.i.d., from (3.1) we have

$$L(\boldsymbol{\theta}; \mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}).$$

We say that $L(\boldsymbol{\theta}; \mathbf{x})$ is the likelihood of $\boldsymbol{\theta}$ based on data $\mathbf{x}$.

The likelihood is a function of the parameter $\boldsymbol{\theta}$. Thus, to completely describe the likelihood, it is important when we calculate $L(\boldsymbol{\theta}; \mathbf{x})$ to identify the set of the possible parameter values $\boldsymbol{\theta}$, that is the domain of L. We will denote the set of possible parameter values by Θ.

[1] It may seem odd to declare that f is unknown, and then assume that in fact f takes a particular form with only unknown parameters. There are statistical methods aimed at handling *completely* unknown f, but they are outside of the scope of this course.

Example 28 *Discrete likelihood*

Example 29 *Chemical reaction revisited*

3.3 Maximum likelihood estimation

3.3.1 Introduction

The likelihood function $L(\boldsymbol{\theta};\mathbf{x})$, based on the sample data $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, is a function of $\boldsymbol{\theta}$. The meaning of $L(\boldsymbol{\theta};\mathbf{x})$ is that its value gives a measure of "how likely" it is that $\boldsymbol{\theta}$ gives the true values of the parameter of interest, given the random sample $\mathbf{x}$.

If we take two different values of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, then a question arises on which of the two we should choose, given the sample $\mathbf{x}$. The above might suggest that if $L(\boldsymbol{\theta}_1;\mathbf{x}) \geq L(\boldsymbol{\theta}_2;\mathbf{x})$, then $\boldsymbol{\theta}_1$ should be preferred, because $\boldsymbol{\theta}_1$ is "more likely" to describe the generating process of the underlying experiment. This idea leads to the principle of maximum likelihood estimation:

We estimate $\boldsymbol{\theta}$ by the value $\widehat{\boldsymbol{\theta}}$ which maximises the likelihood function, i.e.

$$\widehat{\boldsymbol{\theta}} \text{ is chosen so as } L(\widehat{\boldsymbol{\theta}};\mathbf{x}) \geq L(\boldsymbol{\theta};\mathbf{x}), \text{ for all } \boldsymbol{\theta} \in \Theta,$$

where Θ is the set of all possible parameter values $\boldsymbol{\theta}$. The maximizing value $\widehat{\boldsymbol{\theta}}$, if one exists, is called the **maximum likelihood estimate** of $\boldsymbol{\theta}$, given the data $\mathbf{x}$. If we were given different data, it would typically give us a different maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$.

In some cases $\boldsymbol{\theta}$ will consist of just one parameter, in which case we say we have a **one-parameter problem**, and in some cases $\boldsymbol{\theta}$ will consist of two or more parameters, in which case we say we have a **multi-parameter problem**. In the former case we can write $\boldsymbol{\theta} = \theta$ (a scalar parameter) and we will want to maximise $L(\theta;\mathbf{x})$ over θ. In the latter case we will want to maximise $L(\boldsymbol{\theta};\mathbf{x})$ over $\boldsymbol{\theta}$, a multi-dimensional maximisation problem.

Example 30 *Discrete maximisation of likelihood*

Example 31 *Exponential maximum likelihood*

Example 32 *Binomial maximum likelihood*

3.3.2 Maximising the log-likelihood

Maximum likelihood estimation comes down to a maximisation problem. Whether this is easy or difficult depends on (a) the statistical model we use in the form $f(\mathbf{x}|\boldsymbol{\theta})$ and (b) the parameter vector $\boldsymbol{\theta}$. One-parameter problems are clearly easier to handle and in many cases multi-parameter problems require the use of numerical maximisation techniques.

In maximising $L(\boldsymbol{\theta};\mathbf{x})$ it is usually easier to work with the logarithm of the likelihood instead of the likelihood itself. We call the logarithm of the likelihood the **log-likelihood function** and we write

$$\ell(\boldsymbol{\theta};\mathbf{x}) = \log L(\boldsymbol{\theta};\mathbf{x}).$$

Maximising $\ell(\boldsymbol{\theta};\mathbf{x})$ over $\boldsymbol{\theta}$ produces the same estimator $\widehat{\theta}$ as maximising the likelihood, because the logarithm is increasing, i.e.

$$\widehat{\boldsymbol{\theta}} \text{ is such that } \ell(\widehat{\boldsymbol{\theta}};\mathbf{x}) \geq \ell(\boldsymbol{\theta};\mathbf{x}), \text{ for all } \boldsymbol{\theta} \in \Theta.$$

Note that in this course we will be working with natural logarithms, which work well when dealing with the many standard distributions whose p.d.f.s include an exponential term.

3.3.3 The parameter set and maximisation techniques

When we maximise $\ell(\boldsymbol{\theta}; \mathbf{x})$, we need to be careful with the parameter set Θ. In most of the examples we will meet in this module $\boldsymbol{\theta}$ will be continuous and so we can use differentiation to obtain the maximum. However, in some cases, such as Example 30, the possible values of $\boldsymbol{\theta}$ may be discrete (i.e. Θ is a discrete set) and in such cases we cannot use differentiation.

Remark 3.3.1 *Note that saying $\boldsymbol{\theta} \mapsto L(\boldsymbol{\theta}, \boldsymbol{x})$ is continuous is not the same thing as saying that the distribution of $\mathbf{X}$ is continuous!*

3.3.4 One parameter problems

One-parameter problems can be easily handled using the maximisation and minimisation techniques from single variable calculus theory. For example to obtain the maximum of $\ell(\theta; \mathbf{x})$, we first find the solution of

$$\frac{d\ell(\theta; \mathbf{x})}{d\theta} = 0 \Rightarrow \theta = \widehat{\theta} \tag{3.2}$$

and then we check that

$$\left.\frac{d^2\ell(\theta; \mathbf{x})}{d\theta^2}\right|_{\theta=\widehat{\theta}} < 0. \tag{3.3}$$

Note that (3.2) only does not guarantee that $\widehat{\theta}$ is a maximum; it is necessary to check with (3.3). (In some cases, the maximum likelihood estimate of θ does not exist.)

Example 33 *Chemical reaction again*

Example 34 *Poisson maximum likelihood estimation*

Example 35 *Uniform maximum likelihood estimation*

3.3.5 Multi-parameter problems

For multi-parameter problems, where $\boldsymbol{\theta}$ is a vector, a similar procedure can be followed. Here for simplicity we consider only the case where there are 2 parameters (so that $\boldsymbol{\theta}$ is a 2×1 vector) and write $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. Now we find a stationary point $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2)^T$ of the log-likelihood by solving

$$\frac{\partial\ell(\boldsymbol{\theta}, \mathbf{x})}{\partial\theta_1} = 0, \quad \frac{\partial\ell(\boldsymbol{\theta}, \mathbf{x})}{\partial\theta_2} = 0 \Rightarrow \theta_2 = \widehat{\theta}_2. \tag{3.4}$$

Equation (3.4) is the analogue in the two parameter case of equation (3.2) in the one parameter case. The candidate $\widehat{\boldsymbol{\theta}}$ may be a maximum or not, and we have to check this by using an analogue of equation (3.3) in order to check if $\widehat{\boldsymbol{\theta}}$ is indeed a (local) maximum of the log likelihood function. First we calculate the so called **Hessian matrix**:

$$H = \begin{pmatrix} \partial^2\ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_1^2 & \partial^2\ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_1\partial\theta_2 \\ \partial^2\ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_1\partial\theta_2 & \partial^2\ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_2^2 \end{pmatrix}$$

and then we evaluate H at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, where $\widehat{\boldsymbol{\theta}}$ is the stationary point we found using (3.4). If H is a negative definite matrix (the analogue of the second derivative being negative in the one parameter case), then $\widehat{\boldsymbol{\theta}}$ maximises $\ell(\boldsymbol{\theta}; \mathbf{x})$; if H is not a negative definite matrix, then we cannot conclude that $\widehat{\boldsymbol{\theta}}$ is a (local) maximum.

To check that H is a negative definite matrix we can use the following (in the 2 variable case): if

$$\partial^2 \ell(\boldsymbol{\theta}; \mathbf{x})/\partial\theta_1^2 < 0$$

and the determinant $\det(H)$ is positive, then H is negative definite.

(More detail on maximising and minimising functions of more than one variable can be found in the module MAS211 Advanced Calculus and Linear Algebra.)

Example 36 *Maximum likelihood estimation for normal distribution with unknown mean and variance*

3.4 Likelihood regions

Maximum likelihood estimation gives us a single value for the unknown parameters $\boldsymbol{\theta}$, a so-called point estimate. In many settings in statistical inference we want to go further than point estimation, in particular to give some idea of the uncertainty in our point estimate. For example, where we are trying to estimate a single parameter θ, we may want to produce an interval estimate, typically a set of values $[\theta_1, \theta_2]$ which we believe that the true value θ lies in. Alternatively, we may want to test a hypothesis about θ. The likelihood function can often be used to construct appropriate methods in these settings too, and as with maximum likelihood estimation it can often be shown that they are in some sense optimal.

We will start off by thinking about interval estimation. Assume, in the one parameter case, that we have a likelihood function $L(\theta; \mathbf{x})$ defined for $\theta \in \Theta$, maximised at its maximum likelihood estimate $\hat{\theta}$. Then a natural choice of interval estimate is to set some threshold, L_0 say, and to use the values of θ such that $L(\theta; \mathbf{x}) \geq L_0$ as an interval estimate. One natural choice for the threshold is to choose L_0 to be a fixed multiple of the maximum likelihood, say

$$L_0 = e^{-k} L(\hat{\theta}; \mathbf{x})$$

for some chosen $k > 0$. Equivalently in terms of the log-likelihood,

$$\log L_0 = \ell(\hat{\theta}; \mathbf{x}) - k.$$

Our choice of k here will involve a trade off between a precise answer (meaning a narrow interval) and minimising the risk of missing the true value from the interval: a small k will give a narrow interval but relatively low confidence that the interval contains the true value, while a large k will give a larger interval and higher confidence.

More generally, we can make the following definition. The **k-unit likelihood region** for parameters $\boldsymbol{\theta}$ based on data $\mathbf{x}$ is the region

$$R_k = \left\{ \boldsymbol{\theta} : \ell(\boldsymbol{\theta}; \mathbf{x}) \geq \ell(\hat{\boldsymbol{\theta}}; \mathbf{x}) - k \right\},$$

or equivalently

$$R_k = \left\{ \boldsymbol{\theta} : L(\boldsymbol{\theta}; \mathbf{x}) \geq e^{-k} L(\hat{\boldsymbol{\theta}}; \mathbf{x}) \right\},$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ based on $\mathbf{x}$.

The values of $\boldsymbol{\theta}$ within the k-unit likelihood region are those whose likelihood is at least within a factor e^{-k} of the maximum. For instance, points in the 1-unit region have likelihoods within a factor $e^{-1} = 0.368$ of the maximum. The 2-unit region contains points with likelihoods within a factor $e^{-2} = 0.135$ of the maximum. The 2-unit region is the most commonly used in practice.

Example 37 *Interval estimation based on likelihood for normal distributions*

3.5 Hypothesis tests

If we are trying to test a null hypothesis $H_0 : \theta = \theta_0$ against a general alternative hypothesis $H_1 : \theta \neq \theta_0$, then we can use a similar idea: we compare the likelihood of the null hypothesis value, $L(\theta_0; \mathbf{x})$, with the maximum likelihood, $L(\hat{\theta}; \mathbf{x})$ where $\hat{\theta}$ is the maximum likelihood estimate. If the former is much smaller than the latter, then we reject the null hypothesis; if not we do not. We can make this precise by saying that we reject the null hypothesis if and only if the ratio

$$\frac{L(\theta_0; \mathbf{x})}{L(\hat{\theta}; \mathbf{x})} \leq e^{-k},$$

where the parameter k is chosen in a similar way to when finding a likelihood region.

Example 38 *Hypothesis tests based on likelihood for normal distributions*

This leads into the idea of **likelihood ratio tests**, which you will see more of if you take further courses in statistics.

Chapter 4

Likelihood: Case Studies

4.1 Animal study

A vet proposes a new treatment protocol for a certain animal disease. With current methods about 40% of animals with this condition survive six months after diagnosis.

- After one year of using the new protocol, 15 animals have been treated and followed to six months after diagnosis, of whom 6 survived. This constitutes our first set of data.
- After two years a further 55 animals have been followed to the six months mark, of whom 28 survived. Combining with the case above, in total we have 34 animals surviving out of 70. This constitutes our second set of data.

For completeness, we note that this case study (unlike the next two) is not using real data, but it is using data based on real examples. Such data can arise in preliminary investigations of new treatments. Promising results (such as our first data set) might well be followed up by a larger, more carefully controlled trial (such as our second data set).

We model our data as independent Bernoulli trials, giving the binomial observation $X =$ number of animals surviving, and $X \sim Bi(70, \theta)$, where θ is the probability of survival with the new treatment. We have actually observed $x = 34$.

Interest in this case study clearly rests directly on whether the new protocol increases the six-month survival rate, i.e. whether $\theta > 0.4$.

By also looking at the results after the first year, we can watch how the evidence might build up over time. After the first year, we have $y = 6$ survivors, modelled as $Y \sim Bi(15, \theta)$.

Example 39 *Likelihood calculations for animal study*

4.2 Cost effectiveness

The framework

An area of growing importance is that of judging the cost-effectiveness of medical treatments. Very expensive treatments cannot be justified even when they are very effective, since there is only a finite total resource to devote to health care.

These data concern the comparison of two treatments for depression (42 on Antidepressants and 39 on Counselling) for efficacy and for cost. It is a simplification of a rather larger collection

(103 patients) with more information on illness: boiled down to good and not good (those without enough on record to do this were omitted, which is a potential bias). It isn't clear what exactly the costs related to.

The Data

There are data on 81 patients, 42 on Antidepressants and 39 on Counselling; outcome 'good' or 'not good' (after a year); costs (of treatment) over the year are in £. The interest will be in comparing mean costs: hence the question is how to estimate this well. In this course we will focus on the costs for those on antidepressants only.

The data are those used in:
Stevens J., O'Hagan A. and Miller P. (2003). Case study in the Bayesian analysis of a cost-effectiveness trial in the evaluation of health care technologies: Depression. *Pharmaceutical Statistics* **2**, 51–68.

A model for those on antidepressants

Based on the histograms (figures 4.1 and 4.2, produced in R), it would not be reasonable to suppose that the costs followed a normal distribution. For skewed strictly positive data (like costs) working with the log of the original values is often a good idea, and their histogram looks roughly bell-shaped. Hence, to model the situation we might let $x_i = \log c_i$, $(i = 1, 2, \ldots, 42)$, and suppose that the x_is are independent observations from a normal distribution. So we will assume $X_i \sim N(\mu, \sigma^2)$ – so there are two parameters, μ and σ^2. [We say that $C_i = \exp X_i$ has a **lognormal distribution**, $C_i \sim lN(\mu, \sigma^2)$.]

Example 40 *Likelihood calculations for antidepressant costs*

4.3 Ecotoxicology

The framework

These data arose in a real problem concerning standards for toxic chemicals in rivers. The purpose of setting a standard is to control the concentrations of pollutants at levels that protect aquatic animals. However, there are a very large number of species that there is a need to protect, not just fish but also water snails, insects, leeches, etc. Data on toxicity of any given chemical is available for only a small number of the species that are of interest.

In this analysis, all the available toxicity data for the insecticide chlorpyrifos was identified. The data are estimates of the LC50 (the concentration that will kill 50% of the individuals) for 96-hour exposure to chlorpyrifos. The 11 items of 96-hour LC50 data found are tabulated. There is a wish to extrapolate from the published data to say something about the wider collection of species that will be present in a river. To augment the data, some freshwater biologists score the sensitivity to chlorpyrifos of each of 96 taxonomic groups on a scale from 1 (insensitive) to 8 (highly sensitive). They were also asked to score their own knowledge of each taxonomic group from 0 (no knowledge) to 5 (high level of knowledge). The sensitivity scores for each taxon were then weighted according to expertise. In the data, 11 taxonomic groups are represented.

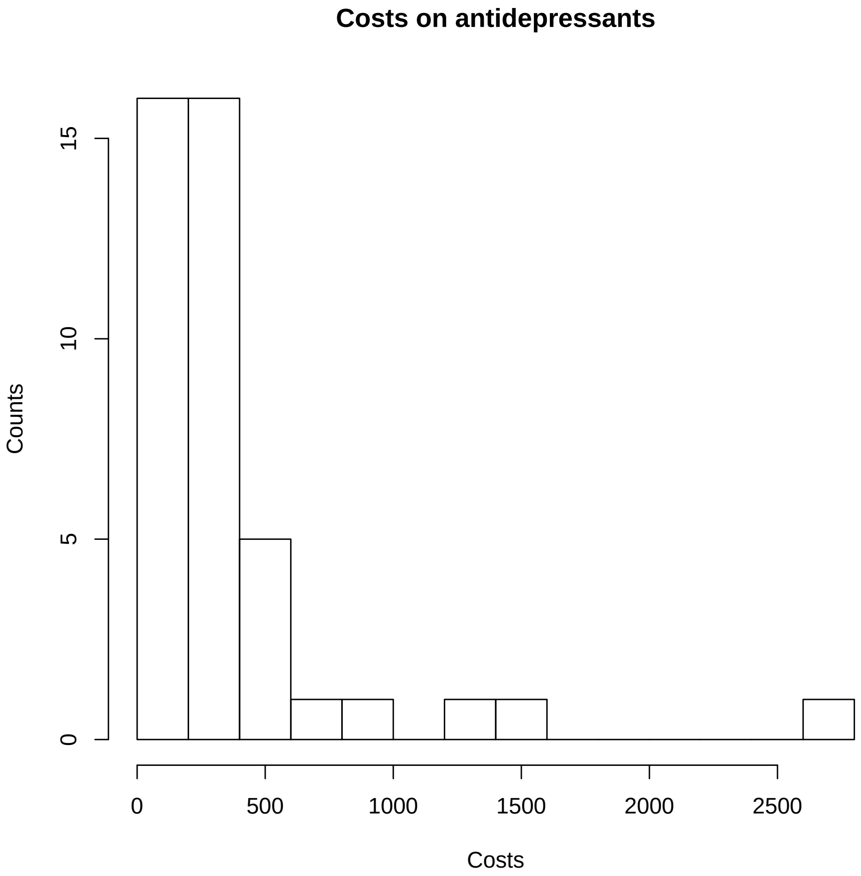

Figure 4.1: Histogram of Antidepressant costs. The mean cost is £379 and the sample variance is 2.25×10^5.

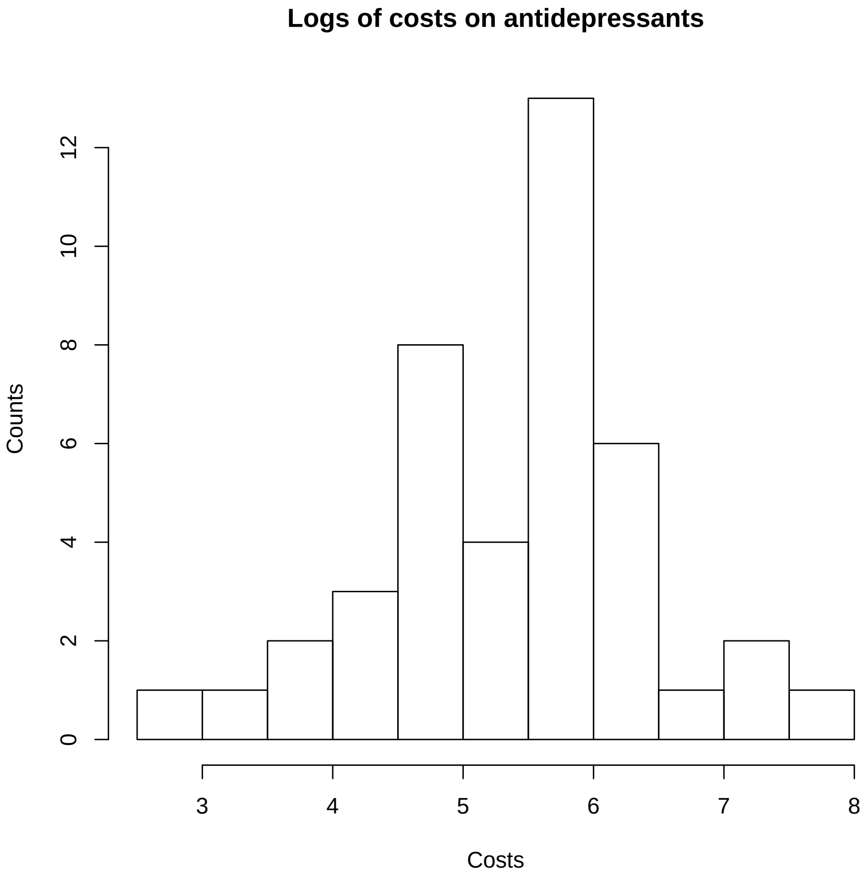

Figure 4.2: Histogram of Antidepressant log-costs. The mean is 5.45 and the sample variance is 1.02.

The Data

Species	*Taxon*	*96hr LC50 ($\mu g/L$)*	*Expert*
Anguilla anguilla	Anguillidae (eels)	540	4.17
Asellus aquaticus	Asellidae (water hoglice)	2.7	4.08
Caenis horaria	Caenidae (mayflies)	0.5	5.86
Chironomus tentatus	Chironomidae (midges)	0.47	3.56
Corixa punctata	Corixidae (lesser waterboatmen)	2	5.11
Rutilus rutilus	Cyprinidae (carp)	120	4.08
Gammarus lacustris	Gammaridae (shrimps)	0.11	5.57
Pungitus pungitus	Gasterosteidae (sticklebacks)	4.7	4.13
Peltodytes sp.	Haliplidae (water beetles)	0.8	5.00
Leptocerida sp.	Leptoceridae (caddis flies)	0.77	6.00
Oncorhynchus mykiss	Salmonidae (salmon)	7.1	5.40

Asides

- The technical term for a taxonomic group is a *taxon*, plural *taxa*.
- The idea of getting the expert assessments was to try to link these to the toxicity data, in order to predict toxicity in other species. It is not practicable to ask the experts about all the possible species that might be present in a British river, so they were asked about groups (taxa) that contain related species (which hopefully will have similar sensitivity to chlorpyrifos).
- The original analysis, which used more data and had more complications, can be found in

  Grist, E.P.M., O'Hagan, A., Crane, M., Sorokin, N., Sims, I. and Whitehouse, P. (2006). Bayesian and time-independent species sensitivity distributions for risk assessment of chemicals. *Environmental Science and Technology* **40**, 395–401.

A model

The following statistical model will be assumed for these data. For $i = 1, 2, \ldots, 11$, let $y_i = \log z_i$, where z_i is the ith toxicity measurement, and let x_i be the corresponding average score of the experts. Suppose that a linear regression relationship applies between these variables, so that

$$y_i = \alpha + \beta x_i + \epsilon_i \ ,$$

where the ϵ_is are independent $N(0, \sigma^2)$ errors, and $\boldsymbol{\theta} = (\alpha, \beta, \sigma^2)$ are the unknown parameters. We should expect that β will be negative, because increasing sensitivity to chlorpyrifos should be associated with a lower LC50, and Figure 4.3 provides support to this expectation.

Example 41 *Setting up likelihood for linear model*

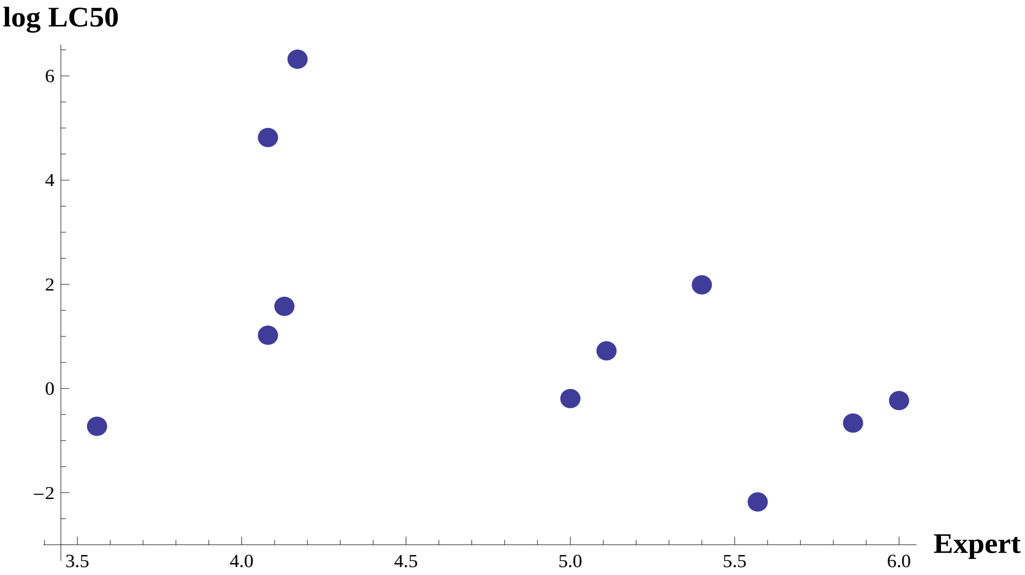

Figure 4.3: Toxicity data. The log of LC50 plotted against expert sensitivity scores appears to support a linear relation between them.